心理学报 2019, Vol. 51, No.12, 1386−1397
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2019.01386

基于分部评分模型思路的多级评分认知诊断模型开发*

高旭亮[1,2] 汪大勋[1] 王 芳[2] 蔡 艳[1] 涂冬波[1]

([1] 江西师范大学心理学院，南昌 330022) ([2] 贵州师范大学心理学院，贵阳 550000)

摘 要 基于分部评分模型的思路，本文提出了一般化的分部评分认知诊断模型(General Partial Credit Diagnostic Model, GPCDM)，与国际上已有的基于分部评分模型思路的多级评分模型 GDM (von Davier, 2008)和 PC-DINA (de la Torre, 2012)相比，GPCDM 的 **Q** 矩阵定义更加灵活，项目参数的约束条件更少。Monte Carlo 实验研究表明，GPCDM 模型的参数估计精度指标 RMSE 介于[0.015, 0.043]，表明估计精度尚可；TIMSS (2007)实证数据应用研究表明，与 GDM 和 PC-DINA 模型相比，GPCDM 与该数据的拟合度更好，并且使用 GPCDM 分析该数据的诊断效果也更优。总之，本研究提供了一种约束条件更少、功能更为强大的多级评分认知诊断模型。

关键词 认知诊断；多级评分认知诊断模型；GDM；PC-DINA

分类号 B841

1 引言

目前，教育评估和心理计量学的最新发展越来越强调形成性评估(Formative Assessments)，它可以提供更多的信息来改进学习和教学策略。认知诊断评估(Cognitively Diagnostic Assessments, CDA)旨在测量特定的知识结构和加工技能，从而为教师和学生提供即时的诊断信息，以便对课堂教学进行相应的规划或修改，以促进个体的全面发展(de la Torre & Minchen, 2014; Leighton & Gierl, 2007)。特别地，美国 2001 年通过了《不让一个孩子掉队法》法案(No Child Left Behind Act of 2001)，法案要求测验要给学生、家长和老师提供有价值的诊断性报告，报告要包括关于学生在解决问题时所需的基础知识和认知处理技能等方面的掌握信息，从而为学生提供量身定制的教育服务。美国政府 2015 年再次通过了每个学生成功(Every Student Succeeds Act)教育法案，新法案继续强调测验要为学生及家长提供诊断性评价、形成性评价。我国在 2010 年通过的《国家中长期教育改革和发展规划纲要(2010−2020 年)》强调要注重因材施教，减轻学生负担，改革教学评价制度，建立科学的教育质量评价体系等。从国内外的教育政策可见，CDA 在未来的教育评估领域将会发挥更大的作用。

当前，研究者已经开发了大量的二级(0-1)评分认知诊断模型(Cognitive Diagnosis Model, CDM)，然而在实际教育和心理评估测验中存在大量多级评分的数据，例如，心理测验中经常使用李克特型(Likert-type)量表问卷，在态度倾向性的问卷中，使用“完全不同意”, “不同意”, “不确定”, “同意”和“完全同意”等 5 个选项来表示不同程度的态度倾向，每个选项代表不同的得分。不仅如此，与二级评分的题目相比，多级评分题目可以提供更多的信息，它只需要更少的题目就能达到和较多二级题目同样的测量精度(van der Ark, 2001)。

为了分析多级评分数据，一个常用的方法是将

收稿日期: 2019−02−12

* 国家自然科学基金(31660278, 31760288, 31960186)资助。

汪大勋为共同第一作者。

通信作者：涂冬波, E-mail: tudongbo@aliyun.com。

多级评分数据转换为二级评分，然后再使用二级评分的 CDM 来分析(Templin & Henson, 2006)。然而，经过转换之后必然要损失很多有价值的信息，Ma 和 de la Torre (2016)以及 Tu, Zheng, Cai, Gao 和 Wang (2017)的研究均发现，与使用多级评分模型相比，使用二级评分模型分析多级评分数据会在很大程度上降低测验的精度。

Mellenbergh (1995)根据模型将多级评分数据二级化的方式将 IRT 的多级评分模型分为 3 类: (1)累积概率(cumulative probability models)模型，或者也被称作等级反应(graded-response models)模型，它是基于全局或累积 logit (global or cumulative logit)的一类模型；(2)连续比率(continuation ratio models)模型，或者也被称作顺序(sequential)模型，它是基于连续比率 logit (continuation ratio logit)的一类模型；(3)相邻类别(adjacent category)模型，或者也被称作分部评分(partial-credit)模型，它是基于局部或相邻类别 logit (local or adjacent category logit)的一类模型。这 3 类模型将多级评分数据二级化的方式是完全不同的，假设题目满分是 3 分，定义 $t=1,2,3$，累积概率模型(cumulative probability models)二分为 $P(x \ge t)$ 和 $P(x < t)$，而连续比率模型(continuation ratio models)则二分为 $P(x \ge t)$ 和 $P(x = t-1)$，相邻类别模型(adjacent category models)二分为 $P(x = t)$ 和 $P(x = t-1)$。因此，这 3 类模型的建模思路是完全不同的，各有特点，累积概率模型侧重于分析某个等级以上(包括该等级)所有等级与该等级下(不包括该等级)所有等级之间的关系；连续比率模型侧重于分析某个等级以上(包括该等级)与该等级的向下一个等级之间的关系；而相邻类别模型侧重于分析两个相邻类别之间的关系。因此，累积概率模型是从整体出发考虑模型的建构，这类模型更适用于分析不强调具体解题步骤的诊断测验，例如，写作水平测验。而连续比率模型和相邻类别模型都是基于解题步骤(steps)来考虑模型的建构，但连续比率模型更强调作答过程是连续步骤(consecutive steps)，即只有成功地完成前面的所有步骤，才能成功地执行下一步，它适合分析解题步骤之间具有严格顺序关系的题目；而相邻类别模型是基于一个局部步骤(local step)来建模，即被试在当前步骤的解答只和前一步有关，这类模型更适合分析相邻步骤之间具有依赖关系的题目。Tutz (1997)认为相邻类别模型更适合分析评定量表(rating scales)类型的题目，连续比率模型更适合分析解答过程包含一系列连续步骤的题目。

在 CDA 领域，研究者已经开发了少量的多级评分 CDMs (polytomous CDMs)。但是已有的多级评分 CDMs 主要是属于累积概率(cumulative probability)模型和连续比率(continuation ratio)模型。Hansen (2013)借鉴 Samejima (1969)等级反应模型(Graded Response Model, GRM)的思想，提出了多级评分的 LCDM 模型。涂冬波、蔡艳、戴海琦和丁树良(2010)基于等级反应模型(GRM)的建模思路提出了多级评分的 DINA 模型(polytomous DINA, P-DINA)。蔡艳、苗莹和涂冬波 (2016)在 P-DINA 模型的基础上加以改进，提出了拓广的 P-DINA (Generalized P-DINA, GP-DINA)模型。Ma 和 de la Torre (2016)在 G-DINA 模型的基础上提出了序列加工 G-DINA 模型(sequential G-DINA)，序列加工 G-DINA 模型是基于连续比率(continuation ratio)模型的一个特例。

然而，目前对于相邻类别(adjacent category)或者分部评分(partial-credit)类的多级评分 CDMs 的研究还相对薄弱。已有的分部评分多级 CDMs 模型仅有 von Davier (2008)提出的一般诊断模型(General Diagnostic Model, GDM)和 de la Torre (2012)提出的分部评分 DINA (Partial Credit DINA, PC-DINA)模型。但这两个模型具有以下缺陷：

(1) 首先，这两个模型的 **Q** 矩阵均定义在题目水平(item level)，即它们的一个潜在假设是同一题目中不同得分类别考察的属性是相同的，但是，这可能会导致部分诊断信息的丢失。因为，不同得分类别所考察的属性可能是不同的，如果将 **Q** 矩阵定义在类别水平(category level)可以提供更多的诊断信息，从而提高诊断测验的估计精度。为了方便，题目水平(item level)和类别水平(category level)的 **Q** 矩阵分别简称为 Item-**Q** 和 Cat-**Q**。现以一个例子来说明两种 **Q** 矩阵的区别(见表 1)，例如，$\sqrt{8.5/0.5-8}$ 这道数学题目考察了 3 个属性，A1 表示减法；A2 表示除法；A3 表示开平方。Cat-**Q** 第一步考察了 A2 属性，第二步考察了 A1 属性，第三步考察了 A3 属性。而 Item-**Q** 则假设每个得分类别考察的属性等于整个题目考察的属性，即每一步都考察了 A1, A2 和 A3 这 3 个属性。

(2) 其次，对于 GDM 模型而言，它假设属性之间不存在交互效应，即它只考虑了属性的主效应。而在实际的数据中，属性之间常常存在交互效应，即被试答对题目的概率不仅受到属性主效应的影

表 1 两种不同类型的 Q 矩阵示例

步骤	得分类别	Cat-Q			Item-Q		
		A1	A2	A3	A1	A2	A3
		减法	除法	开方	减法	除法	开方
$\sqrt{8.5/0.5-8}$					1	1	1
步骤 1: $8.5/0.5=17$	1	0	1	0			
步骤 2: $17-8=9$	2	1	0	0			
步骤 3: $\sqrt{9}=3$	3	0	0	1			

响，还受到属性之间交互效应的影响；(3)对于 PC-DINA 模型来说，它是基于 DINA 模型而提出的，DINA 模型假设属性没有主效应，仅有所有属性间的交互效应，它属于具有严格理论假设的简单模型，因此，它不具一般性认知诊断模型的优势。

基于此，本研究重点关注基于分部评分模型的建模思路，开发出新的功能更为强大的多级评分认知诊断模型，以弥补当前国际上基于分部评分模型思路的多级评分 CDMs (如 GDM 和 PC-DINA)的不足。新开发的模型不仅将属性定义在得分类别水平(属性的定义更加精细)，而且它以 G-DINA 模型作为加工函数，因此具有一般性认知诊断模型的优势。

2 基于分部评分模型思路的多级评分 CDM 开发

定义 X_j 表示在第 j 题的作答反应，m_j 表示第 j 题的满分，则 $X_j \in \{0,1,\ldots,m_j\}$，用 K 表示测验考察的属性个数，$\boldsymbol{\alpha}_l$ 表示被试的属性掌握模式，$\boldsymbol{\alpha}_l=(\alpha_{l1},\ldots,\alpha_{lk},\ldots,\alpha_{lK})$，如果属性模式为 $\boldsymbol{\alpha}_l$ 的被试掌握了第 k 个属性，则 $\alpha_{lk}=1$，如果未掌握 $\alpha_{lk}=0$。定义 $P(X_j=x\,|\,\boldsymbol{\alpha}_l)$ 表示属性模式为 $\boldsymbol{\alpha}_l$ 的被试在第 j 题恰得 x 分的概率。$\mathbf{q}_{jx}$ 表示第 j 题得分类别 x 考察的属性向量，$\mathbf{q}_{jx}=(q_{jx1},\ldots,q_{jxK})$，如果 $\mathbf{q}_{jx}$ 包含了第 k 个属性，则 $\mathbf{q}_{jxk}=1$，否则 $\mathbf{q}_{jxk}=0$。

基于局部 logit (local logit) 函数的定义，定义了以下一般化的分部评分认知诊断模型(General Partial Credit Diagnostic Model, GPCDM)表达式：

$$g_x\left[P(X_j=x\,|\,\boldsymbol{\alpha}_l)\right]=\log\frac{P(X_j=x\,|\,\boldsymbol{\alpha}_l)}{P(X_j=x-1\,|\,\boldsymbol{\alpha}_l)}=\beta_{jx0}+\boldsymbol{\beta}_{jx}^T\mathbf{h}\left(\mathbf{q}_{jx},\boldsymbol{\alpha}_l\right) \quad (1)$$

其中 $g_x(\cdot)$ 表示链接函数，即局部 logit (local logit) 函数，β_{jx0} 表示截距参数，$\boldsymbol{\beta}_{jx}^T\mathbf{h}(\mathbf{q}_{jx},\boldsymbol{\alpha}_l)$ 表示属性考察向量 $\mathbf{q}_{jx}$ 和属性掌握模式 $\boldsymbol{\alpha}_l$ 的一组线性组合：

$$\boldsymbol{\beta}_{jx}^T\mathbf{h}\left(\mathbf{q}_{jx},\boldsymbol{\alpha}_l\right)=\sum_{u=1}^{K_{jx}}\beta_{jx,u}\left(\alpha_{lu}q_{jx,u}\right)+\sum_{u=v+1}^{K_{jx}}\sum_{v=1}^{K_{jx}-1}\beta_{jx,uv}\left(\alpha_{lu}\alpha_{lv}q_{jx,u}q_{jx,v}\right)+\cdots+\beta_{jx,12\ldots K_{jx}}\prod_{k=1}^{K_{jx}}\alpha_{lk} \quad (2)$$

上述 K_{jx} 表示第 j 题得分类别 x 考察的属性个数，$\beta_{jx,u}$ 表示 α_{lu} 的主效应，即掌握属性 u 对恰得 x 分的贡献值，$\beta_{jx,uv}$ 表示 α_{lu} 和 α_{lv} 的二阶交互效应，即同时掌握属性 u 和 v 对得 x 分的贡献值，$\beta_{jx,12\ldots K_{jx}}$ 表示 α_{l1} 到 $\alpha_{lK_{jx}}$ 的 K 阶交互效应，即掌握所有属性对得 x 分的贡献。

假设题目的满分是 3 分，即有 4 个得分类别(0, 1, 2, 3)，此时，可以得到每个得分类别的答对概率，如下所示：

$$\begin{cases} g_1\left[P(X_j=1\,|\,\boldsymbol{\alpha}_l)\right]=\log\dfrac{P(X_j=1\,|\,\boldsymbol{\alpha}_l)}{P(X_j=0\,|\,\boldsymbol{\alpha}_l)} \\ g_2\left[P(X_j=2\,|\,\boldsymbol{\alpha}_l)\right]=\log\dfrac{P(X_j=2\,|\,\boldsymbol{\alpha}_l)}{P(X_j=1\,|\,\boldsymbol{\alpha}_l)} \\ g_3\left[P(X_j=3\,|\,\boldsymbol{\alpha}_l)\right]=\log\dfrac{P(X_j=3\,|\,\boldsymbol{\alpha}_l)}{P(X_j=2\,|\,\boldsymbol{\alpha}_l)} \\ P(X_j=0\,|\,\boldsymbol{\alpha}_l)+P(X_j=1\,|\,\boldsymbol{\alpha}_l)+P(X_j=2\,|\,\boldsymbol{\alpha}_l)+P(X_j=3\,|\,\boldsymbol{\alpha}_l)=1 \end{cases} \quad (3)$$

化解公式 3 的方程组，可以得到如下公式：

$$\begin{cases} P(X_j=0\,|\,\boldsymbol{\alpha}_l)=\dfrac{1}{\sum_{r=0}^{m_j}\exp\sum_{c=0}^{r}\left[\beta_{jc}+\boldsymbol{\beta}_{jc}^T\mathbf{h}\left(\mathbf{q}_{jc},\boldsymbol{\alpha}_l\right)\right]} \\ P(X_j=1\,|\,\boldsymbol{\alpha}_l)=\dfrac{\exp\sum_{c=0}^{1}\left[\beta_{jc}+\boldsymbol{\beta}_{jc}^T\mathbf{h}\left(\mathbf{q}_{jc},\boldsymbol{\alpha}_l\right)\right]}{\sum_{r=0}^{m_j}\exp\sum_{c=0}^{r}\left[\beta_{jc}+\boldsymbol{\beta}_{jc}^T\mathbf{h}\left(\mathbf{q}_{jc},\boldsymbol{\alpha}_l\right)\right]} \\ P(X_j=2\,|\,\boldsymbol{\alpha}_l)=\dfrac{\exp\sum_{c=0}^{2}\left[\beta_{jc}+\boldsymbol{\beta}_{jc}^T\mathbf{h}\left(\mathbf{q}_{jc},\boldsymbol{\alpha}_l\right)\right]}{\sum_{r=0}^{m_j}\exp\sum_{c=0}^{r}\left[\beta_{jc}+\boldsymbol{\beta}_{jc}^T\mathbf{h}\left(\mathbf{q}_{jc},\boldsymbol{\alpha}_l\right)\right]} \\ P(X_j=3\,|\,\boldsymbol{\alpha}_l)=\dfrac{\exp\sum_{c=0}^{3}\left[\beta_{jc}+\boldsymbol{\beta}_{jc}^T\mathbf{h}\left(\mathbf{q}_{jc},\boldsymbol{\alpha}_l\right)\right]}{\sum_{r=0}^{m_j}\exp\sum_{c=0}^{r}\left[\beta_{jc}+\boldsymbol{\beta}_{jc}^T\mathbf{h}\left(\mathbf{q}_{jc},\boldsymbol{\alpha}_l\right)\right]} \end{cases} \quad (4)$$

通过公式 4, 进一步可以概括出 GPCDM 模型的每个得分类别的一般化公式:

$$P(X_j = x \mid \boldsymbol{\alpha}_l) = \frac{\exp\sum_{c=0}^{x}\left[\beta_{jc} + \boldsymbol{\beta}_{jc}^T \mathbf{h}\left(\mathbf{q}_{jc}, \boldsymbol{\alpha}_l\right)\right]}{\sum_{r=0}^{m_j}\exp\sum_{c=0}^{r}\left[\beta_{jc} + \boldsymbol{\beta}_{jc}^T \mathbf{h}\left(\mathbf{q}_{jc}, \boldsymbol{\alpha}_l\right)\right]} \quad (5)$$

公式 5 满足 $\sum_{c=0}^{0}\left[\beta_{jc} + \boldsymbol{\beta}_{jc}^T \mathbf{h}\left(\mathbf{q}_{jc}, \boldsymbol{\alpha}_l\right)\right] = 0$。

如果将 **Q** 矩阵定义在题目水平, 即使用 Item-**Q** 时, 并且假设属性没有主效应, 仅保留属性间的最高阶交互效应, 则公式(1)可以简化为:

$$\log\frac{P(X_j = x \mid \boldsymbol{\alpha}_l)}{P(X_j = x-1 \mid \boldsymbol{\alpha}_l)} = \beta_{jx0} + \beta_{jx,12\ldots K_{jx}}\prod_{k=1}^{K_{jx}}\alpha_{lk} \quad (6)$$

从公式(6)可以看出, 此时, GPCDM 等价于 PC-DINA 模型, 这两者的参数可以相互转换, $g_{jx} = \beta_{jx0}$, $1 - s_{jx} = \beta_{jx0} + \beta_{j12\ldots K_{jx}}$。

综上, 已有的分部评分 CDMs 都将 **Q** 矩阵定义在题目水平, 而 GPCDM 的 **Q** 矩阵定义更加灵活, 它可以定义在题目水平和得分类别水平; 当 **Q** 矩阵定义在得分类别时, 即 **Q** 矩阵的定义更加精细, 有助于提供更多的诊断信息。在实际应用中, 使用者可以根据自身的需求灵活选择不同类型的 **Q** 矩阵。另外, GDM 和 PC-DINA 的理论假设均比较严苛, 在应用中具有较大的限制。而 GPCDM 的约束条件更少, 因而, 理论上 GPCDM 在实际应用中更加灵活, 更具优势。

3　参数估计

GPCDM 的参数采用 EM 算法来估计, 用 X_{ij} 表示被试 i 在题目 j 的作答反应, 其中, $i = 1,\ldots,I$ 和 $j = 1,\ldots,J$, m_j 表示题目 j 的满分值, $\mathbf{X}_i$ 表示被试 i 的得分向量 $\mathbf{X}_i = (X_{i1},\ldots,X_{ij})$ 。根据局部独立性假设, 可以得到边际对数似然函数:

$$l(x) = \log\prod_{i=1}^{I}\sum_{l=1}^{2^K} L(\mathbf{X}_i \mid \boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l) \quad (7)$$

这里, $L(\mathbf{X}_i \mid \boldsymbol{\alpha}_l)$ 是属性模式在已知作答向量 $\mathbf{X}_i$ 的似然函数, $p(\boldsymbol{\alpha}_l)$ 是属性模式 $\boldsymbol{\alpha}_l$ 的先验信息, $L(\mathbf{X}_i \mid \boldsymbol{\alpha}_l)$ 可以通过下列公式计算:

$$L(\mathbf{X}_i \mid \boldsymbol{\alpha}_l) = \prod_{j=1}^{J}\prod_{x=0}^{m_j} P(X_{ij} = x \mid \boldsymbol{\alpha}_l)^{I(X_{ij}=x)} \quad (8)$$

$X_{ij} = x$ 表示被试 i 在第 j 题的得分, $I(X_{ij} = x)$ 是一个指示性变量。EM 算法在每次迭代中包括两个步骤: 期望步骤(Expectation Step, E-step)和最大化步骤(Maximization Step, M-step)。E 步是计算属性模式为 $\boldsymbol{\alpha}_l$ 的被试在第 j 题上恰得 x 分的人数, 用 R_{ljx} 来表示,

$$R_{ljx} = \sum_{i=1}^{I} I(X_{ij} = x)P(\boldsymbol{\alpha}_l \mid \mathbf{X}_i) \quad (9)$$

这里 $P(\boldsymbol{\alpha}_l \mid \mathbf{X}_i)$ 表示被试 i 在已知作答向量 $\mathbf{X}_i$ 时属性模式属于 $\boldsymbol{\alpha}_l$ 的后验概率, 可以通过下列公式计算:

$$P(\boldsymbol{\alpha}_l \mid \mathbf{X}_i) = \frac{L(\mathbf{X}_i \mid \boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{2^K} L(\mathbf{X}_i \mid \boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)} \quad (10)$$

对于题目 j, M-step 的目的是使目标函数极大化的条件下来估计项目参数, 目标函数见下列公式 11:

$$f = \sum_{l=1}^{2^K}\sum_{x=0}^{m_j} R_{ljx}\log\left[P(X_{ij} = x \mid \boldsymbol{\alpha}_l)\right] \quad (11)$$

本研究的参数估计程序使用 R 软件来编写, 在 R 软件中 optim 函数包含了几种常用的极值优化算法。optim 函数在 R 里的表达式是 optim (par, fn, method), par 代表项目参数初值, fn 代表目标函数, method 可选择的优化算法, 因此, 使用 optim 函数计算极值时只需要输入 par (项目参数初值), 初值可以从均匀分布中随机生成, fn (目标函数)和选择的优化算法即可。

EM 算法每循环一次, 就验证是否达到收敛条件, 如果达到收敛条件, 则迭代停止, 否则, 重复 E 步和 M 步。最后, 通过 EM 算法得到项目参数后, 采用期望后验(Expected a Posteriori, EAP)方法来估计被试参数(属性掌握模式)。

4　实验 1: Monte Carlo 实验研究

实验 1 旨在检验: (1)GPCDM 模型的参数估计精度及其性能; (2)当采用 Cat-**Q** 矩阵生成数据时, 如果采用 Item-**Q** 矩阵分析数据是否会降低参数估计的精度, Item-**Q** 可以从 Cat-**Q** 得到, 例如, 表 2 中的第 1 题得分类别 1 和 2 考察的属性向量分别是(1, 0, 0, 0, 0)和(0, 1, 0, 0, 0), 而 Item-**Q** 中得分类别 1 和 2 考察的属性向量都是(1, 1, 0, 0, 0)。

自变量包括: (1)样本容量(500, 1000, 2000 和 4000)。(2)属性个数(5 个和 7 个); 5 属性和 7 属性的 Cat-**Q** 见表 2 和表 3, 多级评分题目中每个得分类别最多考察 2 个属性, 并且 Cat-**Q** 中每个属性的测量次数都是相同的。另外, 为了提高诊断测验的效果, 5 属性和 7 属性的 Cat-**Q** 分别包含了 5 个和 7 个二级评分的题目, 且这些测验包括了一个完整的可达矩阵(R 阵)。(3)测验长度, 5 属性时包括 20 和 40 题, 7 属性时包括 25 和 50 题, 40 题和 50 题的 Cat-**Q** 与 20 题和 25 题的 Cat-**Q** 是重复关系。为了减少随机误差, 每种条件下重复模拟实验 100 次。

表 2　5 属性的 Cat-Q 矩阵

题目	得分	A1	A2	A3	A4	A5	题目	得分	A1	A2	A3	A4	A5
1	1	1	0	0	0	0	11	1	1	1	0	0	0
1	2	0	1	0	0	0	11	2	0	0	0	0	1
2	1	0	0	1	0	0	12	1	0	1	0	0	0
2	2	0	0	1	1	0	12	2	0	0	0	1	0
3	1	1	0	0	0	1	12	3	0	0	0	0	1
3	2	1	0	0	0	0	13	1	0	0	0	0	1
4	1	0	0	0	0	1	13	2	0	0	0	1	0
4	2	0	0	0	1	1	13	3	0	0	1	0	0
5	1	0	0	1	0	0	14	1	1	0	0	0	0
5	2	0	1	0	1	0	14	2	0	1	0	0	0
6	1	1	1	0	0	0	14	3	0	0	1	0	0
6	2	0	0	1	0	0	15	1	0	0	0	1	0
7	1	0	1	0	0	0	15	2	0	0	0	0	1
7	2	0	1	0	1	0	15	3	1	0	0	0	0
8	1	0	0	0	1	0	16	1	1	0	0	0	0
8	2	1	0	1	0	0	17	1	0	1	0	0	0
9	1	0	0	0	1	1	18	1	0	0	1	0	0
9	2	0	0	1	0	1	19	1	0	0	0	1	0
10	1	0	1	1	0	0	20	1	0	0	0	0	1
10	2	1	0	0	0	0							

表 3　7 属性的 Cat-Q 矩阵

题目	得分	A1	A2	A3	A4	A5	A6	A7	题目	得分	A1	A2	A3	A4	A5	A6	A7
1	1	1	0	0	0	0	0	0	17	1	0	1	1	0	0	0	0
2	1	0	1	0	0	0	0	0	17	2	1	0	0	0	0	0	0
3	1	0	0	1	0	0	0	0	18	1	1	1	0	0	0	0	0
4	1	0	0	0	1	0	0	0	18	2	0	0	0	0	1	0	0
5	1	0	0	0	0	1	0	0	19	1	1	0	0	0	0	0	0
6	1	0	0	0	0	0	1	0	19	2	0	1	0	0	0	0	0
7	1	0	0	0	0	0	0	1	19	3	0	0	1	0	0	0	1
8	1	1	0	0	0	0	0	0	20	1	0	0	0	0	1	0	0
8	2	0	1	0	0	0	0	0	20	2	0	0	0	0	0	0	1
9	1	0	1	1	0	0	0	0	20	3	0	0	0	1	0	0	1
9	2	0	0	1	1	0	0	0	21	1	0	0	1	0	0	0	0
10	1	1	0	0	0	1	0	0	21	2	0	0	0	1	0	0	0
10	2	1	0	0	1	0	0	1	21	3	0	0	0	0	1	1	0
11	1	0	0	0	0	1	0	0	22	1	0	0	0	1	0	0	0
11	2	1	0	0	0	1	0	0	22	2	0	0	0	0	0	0	1
12	1	0	0	0	0	0	1	0	22	3	0	0	0	0	0	1	1
12	2	0	0	0	0	1	0	1	23	1	0	0	0	0	1	0	0
13	1	0	1	0	0	0	0	0	23	2	0	0	0	0	0	1	0
13	2	0	0	1	0	0	1	0	23	3	0	0	0	0	0	1	1
14	1	0	1	0	0	0	0	0	24	1	1	0	0	0	0	1	1
14	2	0	1	0	1	0	0	0	24	2	0	1	0	0	0	0	0
15	1	0	0	0	1	0	0	0	24	3	0	0	0	0	0	1	0
15	2	1	0	1	0	0	0	0	25	1	0	0	1	0	0	0	0
16	1	0	0	0	1	0	1	0	25	2	0	0	0	0	1	0	0
16	2	0	0	1	0	0	1	0	25	3	0	0	0	0	0	0	1

4.1 参数的模拟

4.1.1 被试参数的模拟

样本容量包含 4 个水平, $N = 500, 1000, 2000$ 和 4000。当属性个数是 5 个时, 所有可能的属性掌握模式是 $2^5 = 32$ 种, 被试的属性模式从 32 种模式中随机生成, 同理, 当考察的属性个数等于 7 个时, 被试的属性模式从 $2^7 = 128$ 种可能的模式中随机生成。

4.1.2 题目参数的模拟

题目参数的模拟方法参考了 Ma 和 de la Torre (2016)的做法, $\text{logit}\{g_x[P(X_j = x|\boldsymbol{\alpha}_l = 1)]\}$ 从均匀分布 $U(0.75,1)$ 中随机生成, 而 $\text{logit}\{g_x[P(X_j = x|\boldsymbol{\alpha}_l = 0)]\}$ 从均匀分布 $U(0,0.25)$ 中随机生成, 这里 $\boldsymbol{\alpha}_l = \mathbf{1}$ 表示被试已经掌握了第 j 题得分类别 x 考察的所有属性, 而 $\boldsymbol{\alpha}_l = \mathbf{0}$ 表示被试未掌握得分类别 x 考察的任意一个属性。当属性模式 $\boldsymbol{\alpha}_l$ 掌握的属性个数介于 $\boldsymbol{\alpha}_l = \mathbf{0}$ 和 $\boldsymbol{\alpha}_l = \mathbf{1}$ 之间时, 即 $\boldsymbol{\alpha}_l \notin \{\boldsymbol{\alpha}_l = \mathbf{0}, \boldsymbol{\alpha}_l = \mathbf{1}\}$, 此时, 属性模式 $\boldsymbol{\alpha}_l$ 相对应的概率从以 $\boldsymbol{\alpha}_l = \mathbf{0}$ 和 $\boldsymbol{\alpha}_l = \mathbf{1}$ 所对应概率为两个边界值的均匀分布中随机生成。

为了保证作答概率满足单调递增性, 即掌握的属性越多答对题目的概率越大, 约定如果属性模式 $\boldsymbol{\alpha}_l$ 的被试掌握的题目 j 考察的属性个数多于 $\boldsymbol{\alpha}_{l'}$, 则 $\boldsymbol{\alpha}_l$ 对应的项目反应概率大于 $\boldsymbol{\alpha}_{l'}$。

4.1.3 作答数据的模拟

根据模拟得到的项目参数, 可以计算属性模式为 $\boldsymbol{\alpha}_l$ 的被试在第 j 题恰得 x 分的概率 $P(X_j = x|\boldsymbol{\alpha}_l)$, 而每个得分类别对应的概率已知, 属性掌握模式为 $\boldsymbol{\alpha}_l$ 的被试在第 j 题的作答从对应的分类分布中抽取。假设被试在某一题恰得 t 分($t \in \{0,1,2,3,4\}$)对应的概率是 $\{0.03, 0.08, 0.12, 0.14, 0.63\}$, 则被试在该题的得分从 $t \in \{0,1,2,3,4\}$ 中抽取一个数, 而每个得分被抽取的概率分别是 0.03, 0.08, 0.12, 0.14 和 0.63。

4.2 评价标准

评价标准包括被试参数和项目参数的返真性, 它们的返真性分别用模式判准率(Pattern Match Rate, PMR)和均方根误差指标(Root Mean Square Error, RMSE)来反映(Ma & de la Torre, 2016)。两个指标的计算公式如下:

$$PMR = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N} I^{(r)}(\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i)}{N \times R} \tag{12}$$

其中 $I^{(r)}(\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i)$ 表示第 r 次实验估计的 $\boldsymbol{\alpha}_i$ 和真值 $\hat{\boldsymbol{\alpha}}_i$ 是否完全相同, 如果相等则 $I^{(r)}(\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i) = 1$, 否则 $I^{(r)}(\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i) = 0$, N 和 $R = 100$ 分别表示人数和实验次数。

$$RMSE = \sqrt{\frac{\sum_{r=1}^{R}\sum_{l=1}^{2^K}\sum_{j=1}^{J}\left[P^{(r)}\left(X_j = x|\boldsymbol{\alpha}_l\right) - \hat{P}^{(r)}\left(X_j = x|\boldsymbol{\alpha}_l\right)\right]^2}{J \times 2^K \times R}} \tag{13}$$

其中 $P^{(r)}\left(X_j = x|\boldsymbol{\alpha}_l\right)$ 和 $\hat{P}^{(r)}\left(X_j = x|\boldsymbol{\alpha}_l\right)$ 分别表示第 r 次实验估计的和真实的属性模式 $\boldsymbol{\alpha}_l$ 在第 j 题得分为 x 的概率。PMR 值越大, RMSE 值越小表示估计误差越小, 表明参数估计算法越有效。

4.3 实验结果

表 4 和表 5 分别显示了各种实验条件下的测验 PMR 指标和 RMSE 指标。

需要强调的是, 作答数据是基于类别水平 **Q** 矩阵(Cat-**Q**)生成的。因此, 为了评估参数估计的精度, 主要关注 Cat-**Q** 的结果。从表 4 的结果可见, 属性个数等于 5 且使用 Cat-**Q** 时, 测验长度在 20 题时, 不同样本容量下的 PMR 值都在 0.94 以上, 而当测验长度增加到 40 题时, 不同样本容量下的 PMR 值均在 0.99 以上。当属性个数等于 7 且使用 Cat-**Q** 时, 在测验长度为 25 题时, 不同样本容量下的 PMR 值都在 0.86 以上, 而在测验长度为 50 题时, 不同样本容量下的 PMR 值都在 0.98 以上。

表 4　各种实验条件下被试参数返真性 PMR 值

属性个数	测验长度	**Q** 矩阵的类型	被试样本容量			
			500	1000	2000	4000
5	20	Item-Q	0.931	0.939	0.943	0.951
		Cat-Q	0.942	0.948	0.949	0.954
	40	Item-Q	0.991	0.993	0.995	0.996
		Cat-Q	0.995	0.996	0.998	0.998
7	25	Item-Q	0.818	0.827	0.852	0.858
		Cat-Q	0.864	0.866	0.868	0.872
	50	Item-Q	0.977	0.979	0.981	0.986
		Cat-Q	0.985	0.987	0.989	0.991

表 5 的结果显示，当使用 Cat-**Q** 时，不管属性个数、测验长度和样本容量如何变化，在所有条件下的测验 RMSE 值均在 0.05 以下。随着样本量的增加，RMSE 也随之降低，例如，属性个数等于 5 和测验长度等于 20 时，在样本容量为 500 的条件下，基于 Item-**Q** 和 Cat-**Q** 的 RMSE 值分别是 0.103 和 0.043，同样的条件下，当样本容量增加到 4000 时，基于 Item-**Q** 和 Cat-**Q** 的 RMSE 值分别降低到 0.053 和 0.015。

表 6 显示了在属性个数为 5，样本容量为 1000，测验长度为 20 题时，Cat-**Q** 和 Item-**Q** 条件下每一题的 RMSE 指标，由于其他实验条件下的结果和表 6 有相似的趋势，因此，限于篇幅的原因，只提供了一种条件下的结果。

从表 6 的结果可以发现，由于后 5 题是二级评分的题目，此时 Cat-**Q** 和 Item-**Q** 是等价的，因此 Cat-**Q** 和 Item-**Q** 的 RMSE 值基本相当，而在多级评分的前 15 题中，基于 Cat-**Q** 得到的 RMSE 值始终要小于基于 Item-**Q** 的 RMSE 值，基于 Cat-**Q** 的最大 RMSE 是 0.036。另外，还可以发现，二级评分题目的 RMSE 要略低于多级评分的题目，这是因为，二级评分题目考察的属性个数要少于多级评分题目。这个结果充分表明，EM 算法可以提供精确的参数估计精度，和 Item-**Q** 相比，使用 Cat-**Q** 有助于提供更多有价值的诊断信息，从而提高诊断测验的精度。

从表 4 和表 5 基于 Cat-**Q** 的结果可以发现，当属性个数等于 5 或 7 时，基于 Cat-**Q** 的 PMR 在短测验(20 题和 25 题)时，分别达到了 0.9 和 0.8 以上，而在长测验条件下(40 和 50 题)时，它们的 PMR 值都在 0.95 以上，它们的 RMSE 值均在 0.05 以下。这充分说明本研究提出的模型参数估计算法可以提供稳健、精确的估计精度。

对比基于不同类别 **Q** 矩阵的结果可以发现，在同样的实验条件下，与基于 Cat-**Q** 结果相比，基于 Item-**Q** 导致更低的 PMR 值，和更高的 RMSE 值。这两种 **Q** 矩阵之间的差异尤其在短测验(5 属性时 20 题或 7 属性时 25 题)或被试人数较少(例如 500 人时)的条件下更加明显，例如，当属性个数等于 7，测验长度为 20，被试人数为 500 人时，从表 4 可以看出，使用 Cat-**Q** 时的 PMR 值大约是 0.86，而当使用 Item-**Q** 时的 PMR 值大约是 0.82。而从表 5 可以

表 5 各种实验条件下的项目参数返真性 RMSE 值

属性个数	测验长度	**Q** 矩阵的类型	被试样本容量			
			500	1000	2000	4000
5	20	Item-**Q**	0.103	0.087	0.067	0.053
		Cat-**Q**	0.043	0.028	0.022	0.015
	40	Item-**Q**	0.101	0.086	0.065	0.052
		Cat-**Q**	0.038	0.028	0.019	0.015
7	25	Item-**Q**	0.104	0.092	0.079	0.049
		Cat-**Q**	0.042	0.032	0.020	0.014
	50	Item-**Q**	0.108	0.089	0.070	0.047
		Cat-**Q**	0.038	0.026	0.019	0.014

表 6 当 $K = 5$ 和 $N = 1000$ 时 20 题的 RMSE 值

题目	**Q** 矩阵的类型		题目	**Q** 矩阵的类型	
	Cat-**Q**	Item-**Q**		Cat-**Q**	Item-**Q**
1	0.025	0.095	11	0.025	0.082
2	0.032	0.092	12	0.026	0.088
3	0.033	0.069	13	0.027	0.091
4	0.036	0.081	14	0.029	0.086
5	0.024	0.086	15	0.028	0.088
6	0.034	0.082	16	0.018	0.019
7	0.033	0.083	17	0.021	0.020
8	0.023	0.079	18	0.019	0.019
9	0.034	0.069	19	0.020	0.019
10	0.024	0.084	20	0.020	0.021

发现, 在同样的条件下, 使用 Cat-**Q** 时的 RMSE 值大约是 0.04, 而使用 Item-**Q** 时, 它的 RMSE 值则大约是 0.1。这些结果都表明如果采用 Item-**Q** 来分析 Cat-**Q** 产生的数据确实会降低项目参数和被试参数的估计精度。这个结论启发实际使用者, 在编写多级评分的诊断题目时, 对于 **Q** 矩阵的标定, 应尽量构建基于得分类别的测验 **Q** 矩阵(即 Cat-**Q**), 使用 Cat-**Q** 有利于提供更多的诊断信息, 从而提高诊断的精度。

5　实验 2: 实证数据研究

5.1　研究目的

为了进一步探讨和比较 GPCDM 在实证数据中的效果, 比较了三个基于分部评分模型思路的多级评分认知诊断模型, 即本文新开发的 GPCDM 以及国际上 GDM 和 PC-DINA 模型, 在国际数学与科学趋势研究(Trends in International Mathematics and Science Study, TIMSS) 2007 四年级数学评估测验数据中的表现。TIMSS 是由国际教育成就评价协会(International Association for the Evaluation of Educational Achievement)发起的一个国际大型教育评估项目, 该项目评估的对象是全球 4 年级和 8 年级的数学与科学学业成就。TIMSS 从 1995 年开始第一次测试, 每 4 年举行一次。在 2015 年的 TIMSS 评估测验中, 来自世界各地的 60 多个国家参加了这次测试。

本文分析了 TIMSS (2007)数据的一个子集, 其中包括 823 名学生对 11 个题目涉及 8 个属性的数据。11 个题目中, 有 3 个多级评分题, 8 个二级评分题目, 它的 **Q** 矩阵见表 7。

5.2　评价标准

评价标准包括以下 3 个方面:

(1) **模型和测验数据整体拟合度**: 通过模型拟合指标: −2 倍对数似然(−2 log-likelihood values, −2LL), Akaike 的信息准则(Akaike's information criterion, AIC; Akaike, 1974), 和贝叶斯信息准则(Bayesian Information Criterion, BIC; Schwarz, 1978)等来比较 3 个模型的拟合度。

(2) **两类特殊被试的诊断属性边际概率(Marginal Probability)**: 两类特殊的被试是指测验得 0 分的被试和得满分(即 14 分)的被试, 一般来说, 得 0 分的被试意味着对所考察的属性基本没掌握, 而得满分的考生应该完全掌握了所考察的属性, 因此, 理论上, 得 0 分的被试估计得到的属性边际概率应该很低(接近于 0), 而得满分的被试估计得到属性边际概率应该很高(接近于 1)。属性边际概率的计算公式如下:

$$\hat{p}_{ik} = \sum_{l=1}^{2^K} P(\boldsymbol{\alpha}_l \mid \mathbf{X}_i)\alpha_{lk} \tag{14}$$

$P(\boldsymbol{\alpha}_l \mid \mathbf{X}_i)$ 计算方法可参考公式(10)。

(3) **认知诊断信度分析**: Templin 和 Bradshaw (2013)提出了一种计算 CDM 下属性信度(attribute reliability)的方法, 该方法可以分为以下几步: (1)首先, 使用选定的 CDM 估计每个被试的属性边际概率; (2)根据第一步估计得到的属性边际概率, 构建

表 7　实证数据的 **Q** 矩阵

Item	Cat	A1	A2	A3	A4	A5	A6	A7	A8
1	1	1	1	0	0	0	0	0	0
2	1	0	1	1	0	1	0	0	0
3	1	1	0	0	0	0	1	0	1
3	2	1	0	0	0	0	1	0	1
4	1	0	1	1	0	0	0	0	0
5	1	0	1	1	0	0	0	0	0
6	1	0	1	0	1	0	0	0	0
7	1	0	1	1	0	1	0	0	0
7	2	0	0	0	0	0	0	1	0
8	1	0	1	1	0	1	0	1	0
9	1	0	1	1	1	0	0	0	0
9	2	0	1	1	1	0	0	0	0
10	1	0	1	1	0	0	0	0	0
11	1	1	1	0	0	0	1	0	1

四格列联表，其中的列联表的四个元素可以通过下列公式计算：

$$\begin{cases} P(\alpha_{.k_1}=1,\alpha_{.k_2}=1)=\frac{1}{N}\sum_{i=1}^{N}\hat{p}_{ik}\hat{p}_{ik} \\ P(\alpha_{.k_1}=1,\alpha_{.k_2}=0)=\frac{1}{N}\sum_{i=1}^{N}\hat{p}_{ik}(1-\hat{p}_{ik}) \\ P(\alpha_{.k_1}=0,\alpha_{.k_2}=1)=\frac{1}{N}\sum_{i=1}^{N}(1-\hat{p}_{ik})\hat{p}_{ik} \\ P(\alpha_{.k_1}=0,\alpha_{.k_2}=0)=\frac{1}{N}\sum_{i=1}^{N}(1-\hat{p}_{ik})(1-\hat{p}_{ik}) \end{cases} \quad (15)$$

这里 $\hat{p}_{ik}$ 表示被试 i 在属性 k 的边际概率，可以通过公式(14)计算得到；(3)根据第 2 步构建的列联表，计算四格相关系数，将四格相关系数当作每个属性的信度指标。

5.3 研究结果

5.3.1 模型拟合结果

表 8 显示了 3 个模型的相对拟合指标，结果显示，GDM 和 PC-DINA 这 2 个模型相比而言，在 3 个拟合指标中，GDM 模型的拟合更优。而这 3 个模型相比而言，GPCDM 在 3 个拟合指标的值都是最小的，即与 GDM 和 PC-DINA 模型相比，GPCDM 是相对拟合更好的模型。

表 8 模型相对拟合指标

模型	拟合指标		
	−2LL	AIC	BIC
GDM	10964	11576	13017
PC-DINA	11191	11757	13089
GPCDM	10598	11312	12993

5.3.2 两类特殊被试的属性边际概率

表 9 显示了 3 个模型估计的两类特殊被试的属性边际概率，对于得 0 分被试而言，3 个模型的平均属性边际概率从低到高顺序依次是：GPCDM、GDM 和 PC-DINA 模型。对比 3 个模型的估计结果可以发现，PC-DINA 模型估计的属性边际概率在 8 个属性上都要明显高于 GDM 和 GPCDM，其中属性 A1 的边际概率达到了 0.548，平均属性边际概率达到了 0.375，PC-DINA 模型会高估这些得 0 分被试的属性边际概率。GDM 模型和 GPCDM 估计的属性边际概率都比较低，两者的平均属性边际概率分别是 0.093 和 0.001，但就具体属性而言，GDM 模型在属性 A7 的边际概率达到了 0.278，与 GPCDM 的结果相比，GDM 模型高估了属性 A7 的边际概率。

对于得满分(14 分)的被试而言，3 个模型的平均属性边际概率从高到低顺序依次是：GPCDM、GDM 和 PC-DINA 模型。PC-DINA 模型只有在属性 A2、A3 和 A7 的属性边际概率达到了 0.9 以上，而在其余属性的边际概率都在 0.7 以下，平均属性边际概率只有 0.749；GDM 模型和 GPCDM 的平均属性边际概率分别是 0.881 和 0.975，但与 GPCDM 相比，GDM 模型在属性 A1、A6 和 A8 的边际概率分别是 0.786、0.671 和 0.671，都明显低于 GPCDM 的 0.984、0.998 和 0.998。

总体来看，对于得 0 分和满分的被试，拟合最优的 GPCDM 模型估计的结果是最合理的，其次是 GDM 模型，最后是 PC-DINA 模型。

5.3.3 属性信度分析

表 10 显示了 3 个模型拟合该实证数据时的属

表 9 两类特殊被试的属性边际概率

分数	模型	A1	A2	A3	A4	A5	A6	A7	A8	*Mean*
0	GDM	0.024	0.000	0.001	0.076	0.062	0.150	0.278	0.150	0.093
	PC-DINA	0.548	0.108	0.387	0.204	0.432	0.470	0.382	0.470	0.375
	GPCDM	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.001
14	GDM	0.786	1.000	0.999	0.980	0.971	0.671	0.975	0.671	0.881
	PC-DINA	0.647	0.988	0.934	0.698	0.601	0.609	0.905	0.609	0.749
	GPCDM	0.984	0.981	1.000	1.000	0.839	0.998	1.000	0.998	0.975

表 10 每个模型下的属性信度

模型	A1	A2	A3	A4	A5	A6	A7	A8	*Mean*
GDM	0.844	0.887	0.899	0.946	0.906	0.997	0.914	0.711	0.888
PC-DINA	0.644	0.716	0.827	0.721	0.507	0.529	0.779	0.529	0.656
GPCDM	0.966	0.907	0.881	0.951	0.873	0.973	0.985	0.841	0.922

性信度，表 10 的最后一列表示 8 个属性的平均信度。对于 GDM 模型而言，属性 A8 的信度指标只有 0.710，是相对最低的，而其余 7 个属性的信度指标都在 0.8 以上，属性信度指标的最高的是 A6 属性，达到了 0.997。对于 PC-DINA 模型而言，属性 A5 的信度指标是相对最低，只有 0.507，而属性 A3 的信度指标最高，但也只有 0.827。而 GPCDM 的 8 个属性最低信度指标是 0.841。

总体而言, PC-DINA 模型的 8 个属性的信度指标都要明显低于 GDM 和 GPCDM。而 GDM 和 GPCDM 相比而言, GPCDM 在属性 A1、A2、A4、A7 和 A8 的信度指标也要高于 GDM 模型，即 GPCDM 在 5 个属性的信度要优于 GDM 模型，GPCDM 在剩余属性 A3、A6 和 A7 的信度指标和 GDM 非常接近。从平均属性信度指标来看，GPCDM 的平均属性信度是最高的，其次是 GDM 模型，最后是 PC-DINA，即 GPCDM 分析该实证数据的效果更优。

6 研究结论与讨论展望

6.1 研究结论

本研究开发了一种更为灵活、功能更为强大，且更有理论意义和应用价值的广义多级评分模型，通过模拟研究验证了 GPCDM 的参数估计精度，最后通过一个实证数据比较了 GPCDM 和已有基于分部评分思路的多级评分 CDMs (GDM 和 PC-DINA) 的应用效果，研究结论主要有:

(1) Monte Carlo 实验研究发现，本研究开发的 GPCDM 的属性模式诊断正确率 PMR 在 5 属性时都在 0.9 以上，项目参数的 RMSE 平均不到 0.05，这表明 GPCDM 模型具有较高的参数估计精度。

(2) 当使用 Item-**Q** 拟合 Cat-**Q** 生成的数据时，题目和被试参数的估计精度都会降低。因此，建议研究者在构建多级评分认知诊断的测验 **Q** 矩阵时，应尽量构建基于得分类别的测验 **Q** 矩阵(即 Cat-**Q**)，它能提供更多的诊断信息。

(3) 最后比较了 GPCDM、GDM 和 PC-DINA 模型在 TIMSS (2007)数据的实际应用效果，结果发现 GPCDM 的模型拟合度更优，并且 GPCDM 分析该数据时的效果也更好。这表明新模型在实践应用中具有一定的优势。

6.2 讨论和展望

为使研究的结果不失一般性以及进一步拓展多级评分 CDMs 的相关研究，未来至少还可以在以下几方面展开研究:

(1) 本研究假设属性之间是相互独立的，**Q** 矩阵的标定完全正确，另外，本研究仅采用了 EAP 方法来估计被试参数，并未对其他方法进行对比研究，这些因素都可能会影响本研究的结论。

(2) 同一份测验中，不同的题目可能拟合不同的 CDM，在二级评分的数据中，de la Torre (2011) 应用 Wald 统计检验的方法为每个题目选择不同的 CDM。而在多级评分数据中，如何为每一题选择最适合的多级评分 CDM 也有待进一步研究。

(3) 多级评分的 **Q** 矩阵可以定义在得分类别水平，这有助于提供更多诊断信息，但是这也会增加 **Q** 矩阵标定的工作量。目前，已经有学者开发了一系列辅助 **Q** 矩阵标定的算法，但这些方法只局限于二级评分的模型。未来的研究可以继续探讨多级评分 CDM 中 **Q** 矩阵的标定算法。

(4) 本研究开发的模型假设考生的解题策略只有一种，但在实际应用中，同一道题目经常存在不同的解题策略。如果在诊断测验中考虑了被试解题策略的差异，这也有助于提供更多有价值的信息，从而提高诊断的精度(涂冬波，蔡艳，戴海琦，丁树良，2012)。因此，开发多策略的多级评分 CDM 值得进一步研究。

(5) 已有的 CD-CAT 相关研究，几乎都是基于二级评分的模型而展开，事实上，多级评分 CD-CAT (Polytomous CD-CAT, PCD-CAT)在实际应用中具有更广阔的前景，不仅是因为心理或教育评估测验中存在大量的多级评分数据，更重要的是与二级评分的题目相比，多级评分题目可以提供更多的信息，即多级评分的 CD-CAT 有助于进一步提高测验的效率，未来的研究可以针对 PCD-CAT 的相关算法展开研究。

参 考 文 献

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.

Cai, Y., Miao, Y., & Tu, D. B. (2016). The polytomously scored cognitive diagnosis computerized adaptive testing. *Journal of Psychological Science, 48*(10), 1338–1346.

[蔡艳，苗莹，涂冬波. (2016). 多级评分的认知诊断计算机化适应测验. *心理学报*, *48*(10), 1338–1346.]

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J. (2012). Application of the DINA model framework to enhance assessment and learning. *In Self-directed learning oriented assessments in the Asia-Pacific* (pp. 87–103). Springer, Dordrecht.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*(2), 89–97.

Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis*. Unpublished doctoral dissertation. University of California at Los Angeles.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 27*(2), 3–16.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 253–275.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*(1), 91–100.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 34*(S1), 1–97.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Templin, J. L. & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251–275.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.

Tu, D. B., Cai, Y., Dai, H. Q., & Ding, S. L. (2010). A polytomous cognitive diagnosis model: P-DINA model. *Acta Psychologica Sinica, 42*(10), 1011–1020.

[涂冬波, 蔡艳, 戴海琦, 丁树良. (2010). 一种多级评分的认知诊断模型: P-DINA 模型的开发. *心理学报*, *42*(10), 1011–1020.]

Tu, D. B., Cai, Y., Dai, H. Q., & Ding, S. L. (2012). A new multiple-strategies cognitive diagnosis model: The MSCD method. *Acta Psychologica Sinica, 44*(11), 1547–1553.

[涂冬波, 蔡艳, 戴海琦, 丁树良. (2012). 一种多策略认知诊断方法: MSCD 方法的开发. *心理学报*, *44*(11), 1547–1553.]

Tu, D., Zheng, C., Cai, Y., Gao, X., & Wang, D. (2017). A polytomous model of cognitive diagnostic assessment for graded data. *International Journal of Testing*, *18*(3), 231–252.

Tutz, G. (1997). Sequential models for ordered responses. *In Handbook of modern item response theory* (pp. 139−152). Springer, New York, NY.

van Der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*(3), 273–282.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307.

Development of a Generalized Cognitive Diagnosis Model for polytomous responses based on Partial Credit Model

GAO Xuliang[1,2]; WANG Daxun[1]; WANG Fang[2]; CAI Yan[1]; TU Dongbo[1]

([1] *School of Psychology Jiangxi normal university, Nanchang* 330022, *China*)

([2] *School of Psychology Guizhou normal university, Guiyang* 550000, *China*)

Abstract

Currently, a large number of cognitive diagnosis models (CDMs) have been proposed to satisfy the demands of the cognitively diagnostic assessment. However, most existing CDMs are only suitable for dichotomously scored items. In practice, there are lager polytomously-score items/data in educational and psychological tests. Therefore, it is very necessary to develop CDMs for polytomous data.

Under the item response theory (IRT) framework, the polytomous models can be divided into three categories: (i) the cumulative probability (or graded-response) models, (ii) continuation ratios (or sequential) models, and (iii) the adjacent-category (or partial-credit) models.

At present, several efforts have been made to develop polytomous partial-credit CDMs, including the general diagnostic model (GDM; von Davier, 2008) and the partial credit DINA (PC-DINA; de la Torre, 2012) model. However, the existing polytomous partial-credit CDMs need to be improved in the following aspects: (1) These CDMs do not consider the relationship between attributes and response categories by assuming that all response categories of an item measure the same attributes. This may result in loss of diagnostic information, because different response categories could measure different attributes; (2) More importantly, the PC-DINA is based on reduced DINA model. Therefore, the current polytomous CDMs are established under strong assumptions and do not have the advantages of general cognitive diagnosis model.

The current article proposes a general partial credit diagnostic model (GPCDM) for polytomous responses with less restrictive assumptions. Item parameters of the proposed models can be estimated using the marginal maximum likelihood estimation approach via Expectation Maximization (MMLE/EM) algorithm.

Study 1 aims to examine (1) whether the EM algorithm can accurately estimate the parameters of the proposed models, and (2) whether using item level Q-matrix (referred to as the Item-Q) to analyze data generated by category level Q-matrix (referred to as the Cat-Q) will reduce the accuracy of parameter estimation. Results showed that when using Cat-Q fitting data, the maximum RMSE was less than 0.05. When the number of attributes was equal to 5 or 7, the minimum pattern match rate (PMR) was 0.9 and 0.8, respectively. These results indicated that item and person parameters could be recovered accurately based on the proposed estimation algorithm. In addition, the results also showed that when Item-Q is used to fit the data generated by Cat-Q, the estimation accuracy of both the item and person parameters could be reduced. Therefore, it is suggested that when constructing the polytomously-scored items for cognitively diagnostic assessment, the item writer should try to identify the association between attributes and categories. In the process, more diagnostic information may be extracted, which in turn helps improve the diagnostic accuracy.

The purpose of Study 2 is to apply the proposed model to the TIMSS (2007) fourth-grade mathematics assessment test to demonstrate its application and feasibility and compare with the exiting GDM and PC-DINA model. The results showed that compared with GDM and PC-DINA models, the new model had a better model fit of test-level, higher attribute reliability and better diagnostic effect.

Key words cognitive diagnosis; polytomous CDMs; GDM model; PC-DINA model